JÜLICH
FORSCHUNGSZENTRUM

Overview of Lustre Usage on JUROPA

26 September 2011 | Frank Heckes, FZ Jülich, JSC

Lustre Status

- ***Lustre Status***
- ***Storage Extension***
- ***Fluctuation in Performance***
- ***Lustre Community Test Cluster***

Lustre Status

- ***Environment***
  - *3288 clients*
  - *OSS (SUN/Nehalem, Bull/Westmere), JBODs, DDN SFA10k*
  - *MDS (Bull/Westmere) Emc Clarion CX-240*
- ***Lustre Version 1.8.4, SLES 11 (SP1)***
  - *Very stable, only minor problems*
- ***$HOME on Lustre***
  - *No other technology needed*
  - *Small file systems (4 OST ~ 28 TB), average file size ~ 1 – 2 kb, Total 24 file systems*
  - *Good experience*
  - *Drawback: Datamigration necessary sometimes*

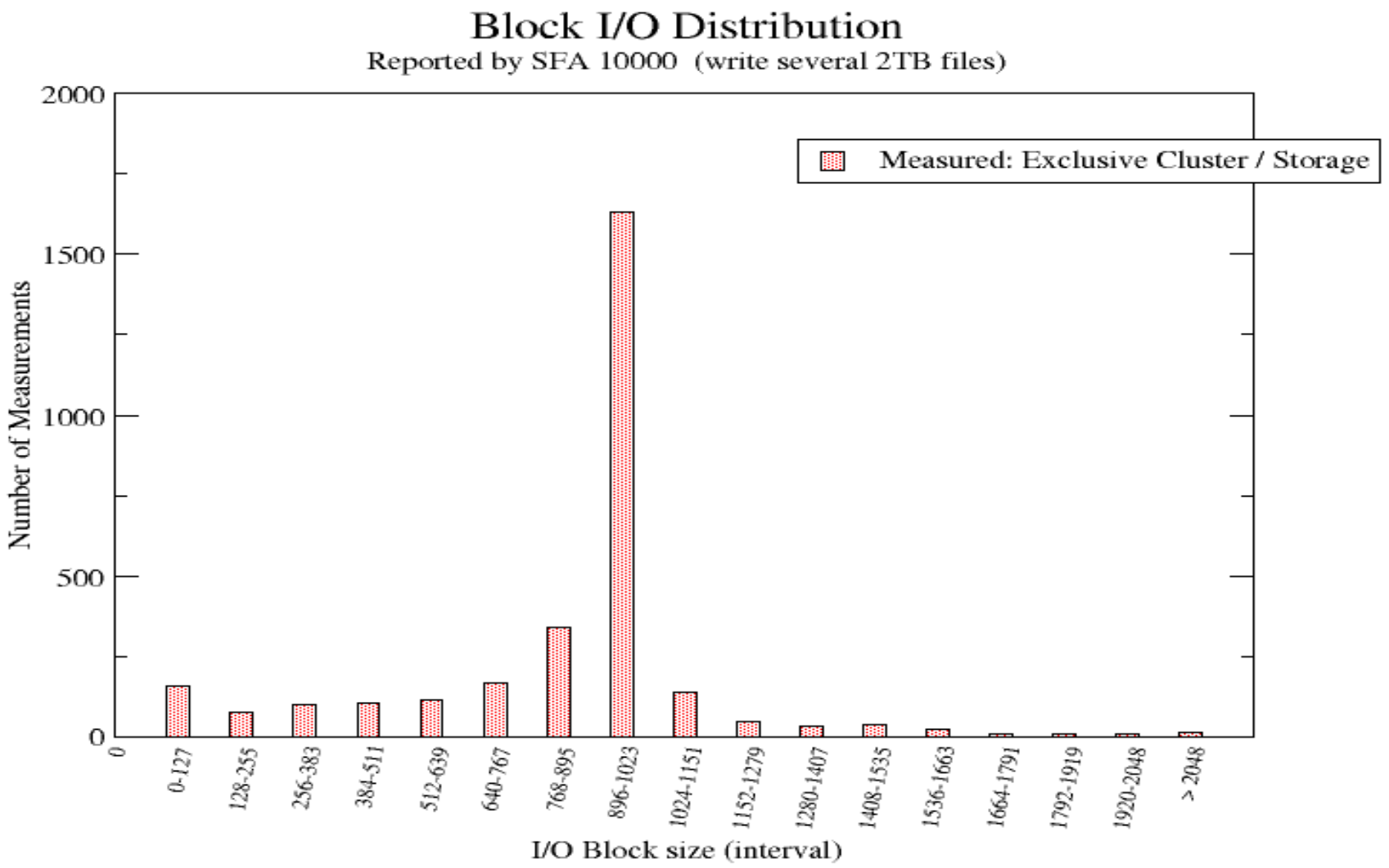
Lustre Status

- ***Bugs***
  - *Sporadic crashing server nodes*
  - *Hangs during server shutdown*
  - *Race condition for clients (fixed in LU-274)*
  - *Problems recursive* `chown/chgrp`
  - *File listing* `ls -color=tty`
  - `mdadm` *re-sync problem*
  - *Many MDT on single MDS ($HOME) might cause performance problems*
- ***Great deviation in Lustre shutdown***
  - *Best values 20 minutes / worst 90*
  - *Needed to reduce downtimes*

Fluctation in Performance

- ***Big deviation in performance***
  - *Test most interesting on scratch file system ($WORK)*
  - *Performance drop: 19.2 GB/s → 14.1 GB/s*
  - *Several reasons*
    - *Fragmented I/O – Lot of read/writes on DDN in range 300 – 1020 kb, even if 1MB blocks are used explicitly*

Fluctuation in Performance



Fluctation in Performance

- ***Big deviation in performance***
  - *Test most interesting on scratch file system ($WORK)*
  - *Performance drop: 19.2 GB/s → 14.1 GB/s*
  - *Several reasons*
    - *Fragmentated I/O – Lot of read/writes on DDN in range 300 – 1020 kb, even if 1MB blocks are used explicitly*
    - *Often not even object distribution for default value of* `qos_threashold_rr` *(0.16).*
    - *Asymmetric allocation of interrupts(?)*
      Handled only by 2 cores; No changes (`smp_affinity`*) possible*
    - *write_throughcache disabled, tuned most common SCSI block parameters* `(max_sectors_kb, nr_requests, timeout,...)`

JÜLICH
FORSCHUNGSZENTRUM

Storage Upgrade

- ***Cluster started with capacity ~900 TB***
  - *Raising number of users and large scale application*
  - *Extend throughput*
- ***Goal: Double amount of storage / throughput and meet acceptance test benchmark***
- ***Upgrade plan***
  - *Replace scratch file system ($WORK) with latest and new hardware*
  - *Re-use parts of previous 'installation' for home directory ($HOME): server, DDN disks, racks → constraints in project schedule*
  - *Additional MDS servers*

JÜLICH
FORSCHUNGSZENTRUM

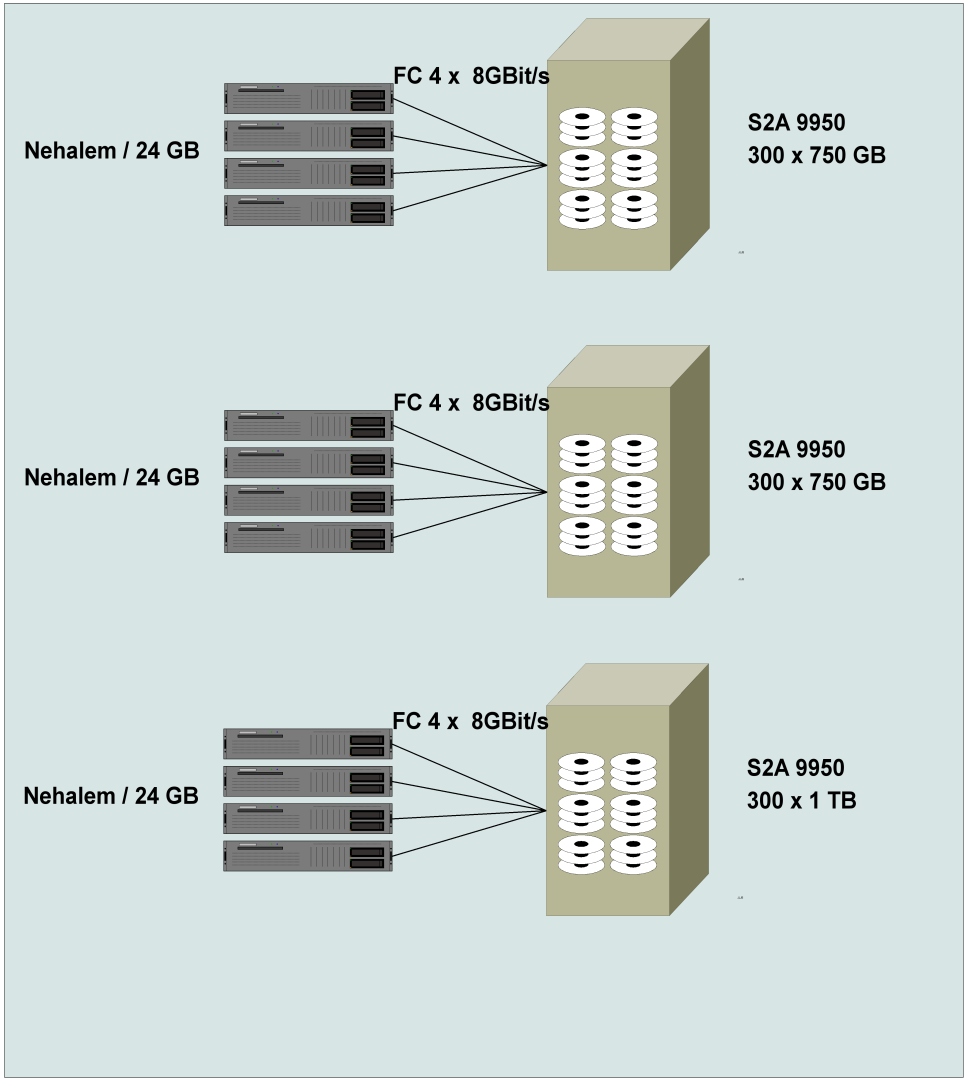
Storage Upgrade

- ***Challenges (before)***
  - *OSS/OST have to be removed from scratch file system*
    - *Lustre Standard migrate procedure went smoothly, but cumbersome*
- ***New scratch file system finished (nearly) on project schedule***
- ***Surprises***
  - *System bus of old server to slow to service four fibre channel interfaces*
  - *A lot of extra benchmarking necessary to drill down problem → several week project delay*
- → ***Use new hardware for home directories, too***

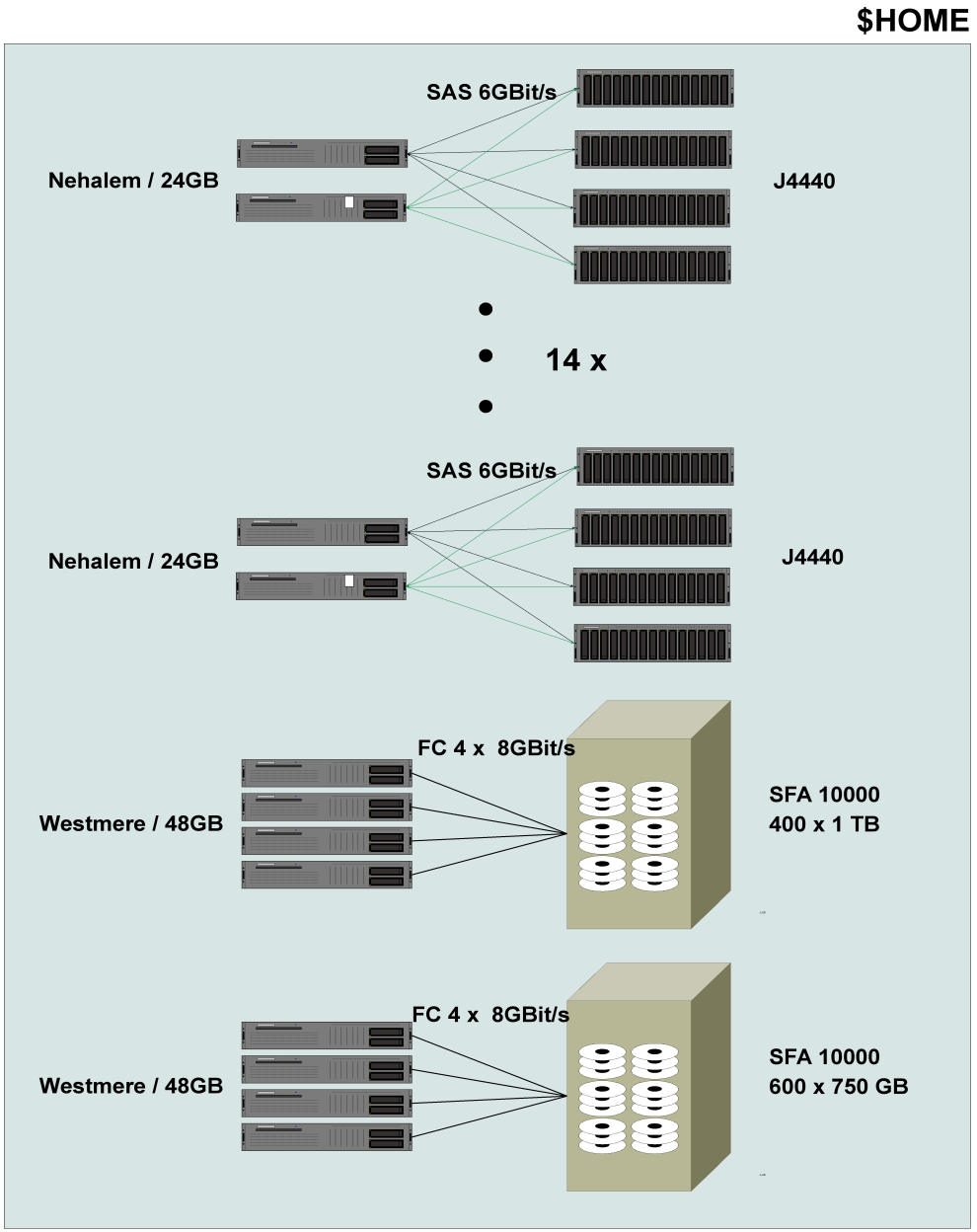
Storage upgrade

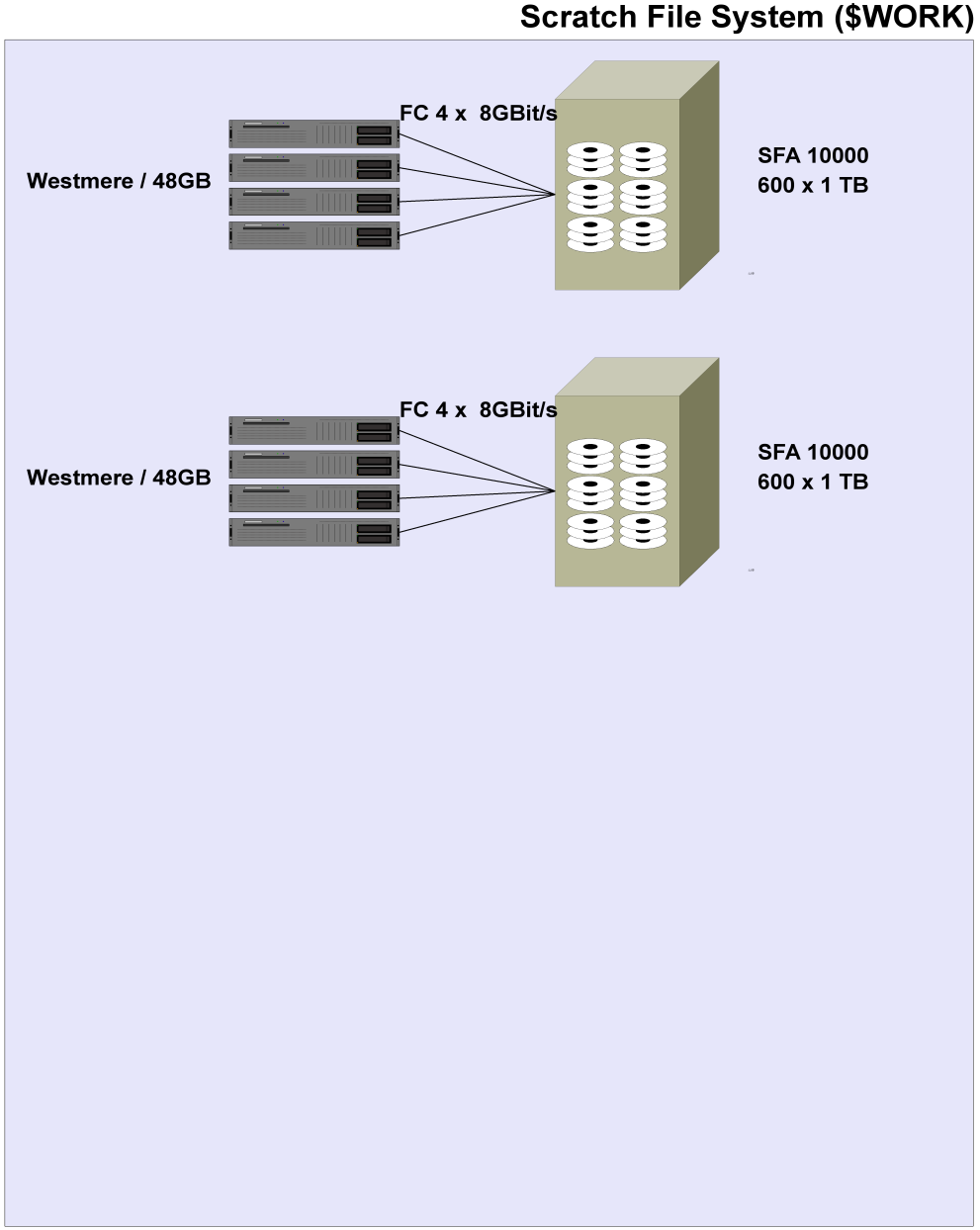
$HOME

Scratch File System ($WORK)



Storage Upgrade

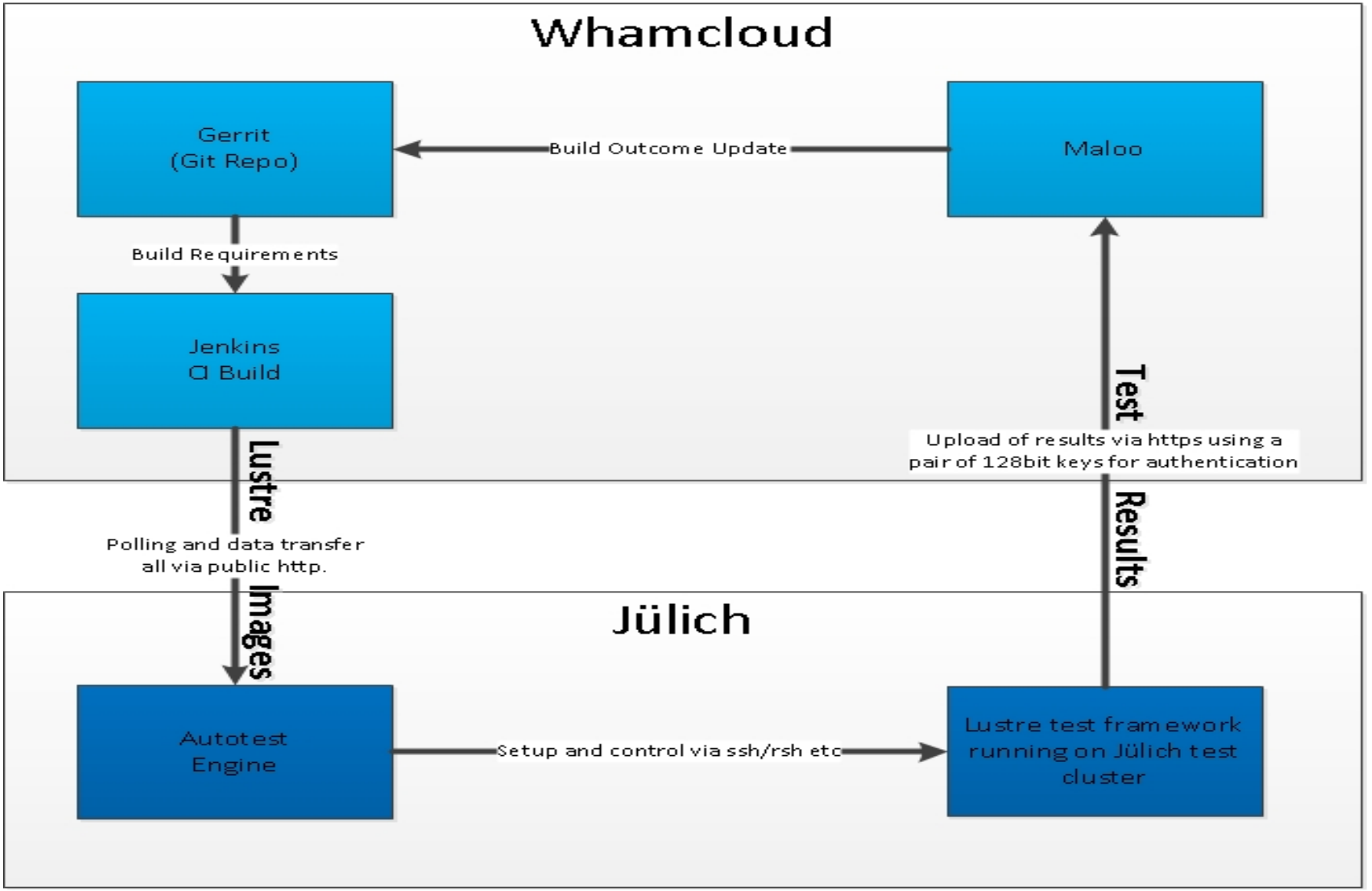




Lustre community test cluster

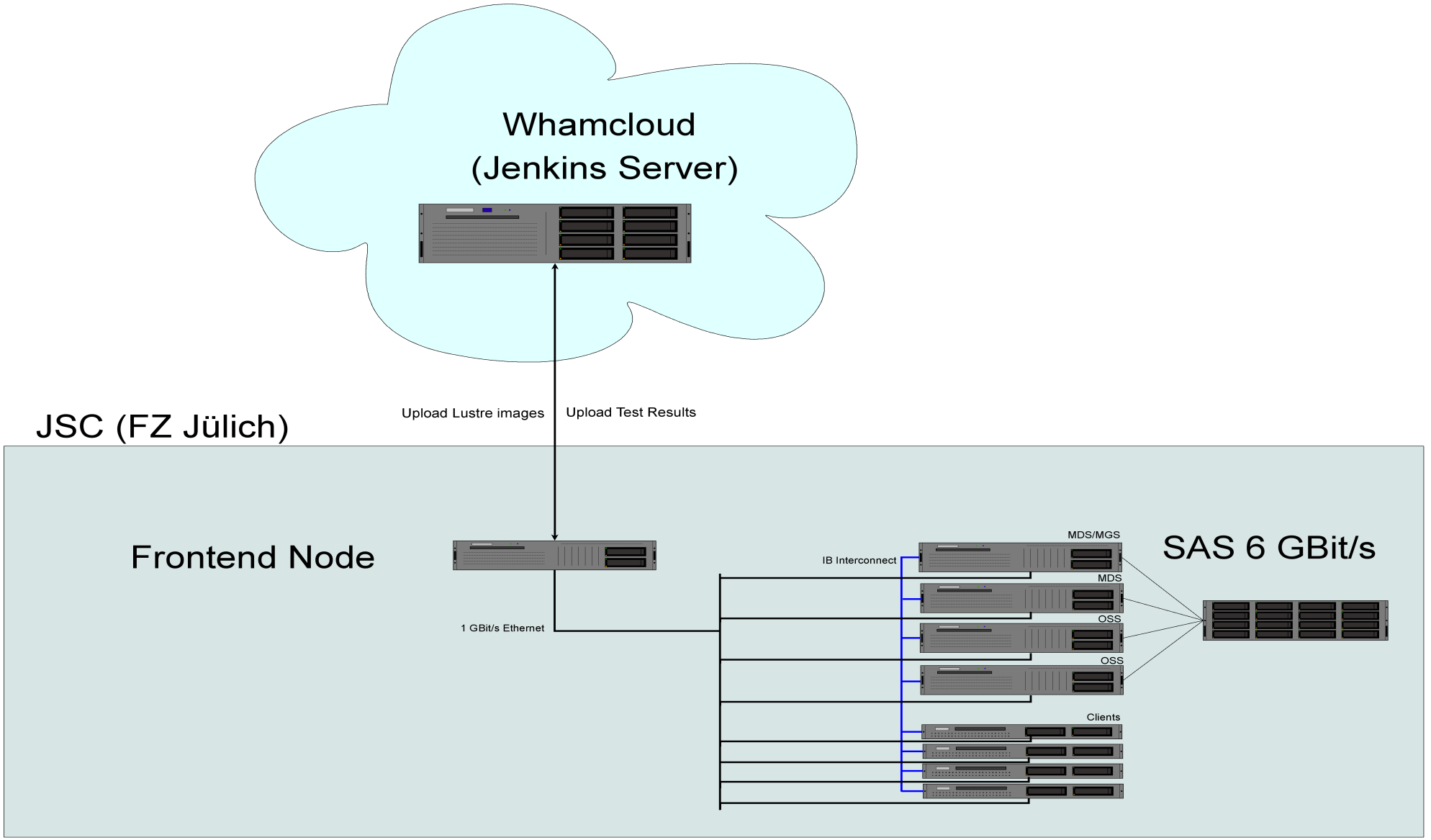
- ***FZJ wants support Lustre development***
- ***Provide test resources for Lustre Small test cluster***
  - *Chance for 'small' sides to contribute*
  - *Cluster rely on automated installation and smoke test framework → minimal administrative overhead*
- ***Hardware Resources***
  - *Frontend node*
  - *2 x OSS, 2 x MDS, 4 x clients*
  - *Enough CPU (Westmere), Memory (24GB) resources for virtualisation*
  - *Infiniband interconnect*
  - *Direct attached storage + SAS switch + software RAID*

Test Cluster (logical view)



By courtesy of Chris Gearing (Whamcloud)

Test Cluster (physical view)



Ongoing Activities

- ***Ongoing projects***
  - *Use `ncheck` command to create file list for client based Tivoli backup*
  - *Implement data mover for IBM Tivoli HSM*
  - *Lustre upgrade >= 1.8.7*
    Download site from Oracle powered down (Oracle support contract)

JÜLICH
FORSCHUNGSZENTRUM

Thank you!